Revitalizing Retail: A Big Data & Analytics Consultancy Approach

Table of Contents

Table of Figures

A. PART-1: Proposal for Daunt Books

1. Introduction

Covid-19 had a significant impact on the book retail industry as people sought escapism through reading(Alexander,2023). Furthermore, recent reports indicate that the book retail industry is expected to grow in the upcoming years despite the challenges posed by inflation and rising living costs(ibid). The strong demand for books is because reading is a cheaper and non-screen entertainment option as well as good for mental well-being (ibid). Although many book readers prefer the shopping experience in physical stores, reports show that multichannel retailers generate more revenue through online than offline(ibid). To gain a competitive edge in the market, many companies are adopting Big Data and Analytics(BDA) technology to formulate business strategies for boosting sales and attracting new customers(Behl *et al.*,2019).

2. Organization Insights

Daunt Books is a reputed medium-sized book retailer based in London, with several outlets across the city including Marylebone, Holland Park, Chelsea, Hampstead, and Belsize Park. Although the major attraction of Daunt Books is travel-literature, it also offers a wide variety of fiction, non-fiction and children's books as well as stationery and gift collections. While the company was forced to close some physical stores during the pandemic, it received increasing demand for books on its website. Thus, the Daunt Book's CEO sensed the need to integrate the channels, enhance marketing strategies and optimise sales. Therefore, the company hired an external consultant to conduct a technical assessment and recommend the best solution using BDA technology.

Through this assessment, the CEO attempts to seek the answer to the following questions:

- How BDA technology can be leveraged to gain customer insights to optimise marketing and sales across online and offline channels?
- How can a dashboard solution be designed for CEO to measure and monitor sales and marketing performance?
- What are the risks involved in the implementation of the proposed system and mitigation steps?

- Provide a high-level implementation plan.

The answers to the above questions will be covered in the upcoming sections. The next section will cover the organisation structure.

2.1. Organisation Structure

The company has dedicated teams for IT, Sales, Marketing, Customer Support, Operations and Supply Chain. The IT team is following the DevOps methodology for delivering software products. Sales and Marketing teams have their own targets and goals including revenue generation, lead generation, conversion rates etc. that need to be achieved at regular time intervals such as monthly, quarterly or annually. The performance of the team is evaluated by the management to ensure that they contribute and align towards the overall organisational goals.

3. Existing System Study

The company uses both traditional and digital marketing methods to attract customers, however, these methods are not implemented effectively enough to exploit their full potential. Fig.1 depicts the pain points experienced by different teams in the existing system. Having understood the pain points of the existing system, the next section will propose a new system to overcome these challenges using BDA technology.

Problem No.	Affected Team	Pain Point	Identified Root Cause
P1	Marketing and Sales	Doesn't have a clear understanding about the customer's shopping behaviour or preferences, making it difficult to launch appropriate marketing campaigns or promotions for the target audience in online and offline sales channels.	Lack of tools or methods to understand customer behaviour or preferences.
P2	Management	Struggles to evaluate the sales and marketing performance without proper analytics tools. This makes it difficult to identify improvement areas, growth opportunities and data-driven decisions.	Lack of tools or methods to evaluate sales and marketing performance.
P3	Supply Chain	Finds it difficult to deliver the product to the customers in different locations on-time due to unavailability of products in the nearest outlets.	Lack of proper inventory management and product distribution strategies in outlets.
P4	Operations	Difficulty in managing the stocks as per the customer demand.	Improper inventory management and sales forecasting.
P5	Customer Service	Received many customer complaints regarding product out-of-stock for longer time than expected. It has encountered scenarios to refund the customer orders due to overselling of books through website. The existing system use SAP for inventory management and the stock level is not updated in real-time on the company website.	Improper inventory management and lack of proper integration between SAP system and website.

Figure 1.Pain Points of Existing System

4. Proposed System using BDA technology

BDA technology is well-suited for the e-commerce industry because bigdata is capable to process huge volumes of diverse data generated through various e-commerce applications and analytics allows businesses to gain meaningful insights from this data using dashboards to make effective marketing and sales strategies (Sayyad *et al.*,2019). The following sections will cover how BDA can be leveraged to overcome the challenges faced by Daunt Books.

4.1. BDA for Enhancing Sales and Marketing Strategies

BDA can be leveraged in various ways to integrate online and offline sales channels to improve marketing effectiveness and optimize sales. The solutions for the challenges identified in the existing system study are covered in the next sections.

4.1.1. Customer Segmentation

The target audience of Daunt Books belongs to different age groups, follow different lifestyles and has varying book genre preferences. Hence, it is important to understand customer requirements and preferences to formulate effective marketing strategies to attract new customers and retain loyal customers.

Benefits of Customer Segmentation in Book Retail Industry	
Customer Retention	Helps to retain online and offline customers by enabling the company to fulfil their changing requirements.
Enhanced Competitiveness	Clear understanding of the target audience preferences gave company a competitive edge over its competitors.
Good Customer Relationships	The company can identify their loyal and premium customers from the huge database that contains information regarding online and offline customers and also allows them to keep a good relationship with them by offering special loyalty services.
Book Price Optimisation	Company can optimise book prices to an extend by understanding the budget of target audience for purchase, resulting in increased sales and conversion rates.
Enhanced Distributed Sales Channel	Helps to identify customer shopping behavior and geographical preference for distributing products to outlets accordingly.
Personalised Marketing Strategies	Helps to tailor marketing strategies that could attract potential customers for repeated purchase.
Increased Revenue	Enables to increase sales revenue by marketing the specific target audience preferences such as specific book genres or formats.
Branding	Customised marketing promotes the brand of the company among online and offline customers.

Figure 2.Benefits of Customer Segmentation in the Book Retail Industry(Bython,2023)

Customer Segmentation provides insights into rapidly changing customer requirements by classifying customers based on their common characteristics and shopping behaviours(Clark,2023;Chiang and Yang,2018). Segmentation will help the marketing team to identify the potential customer behaviour such as their preferred

genre and book format, demographics, location and shopping frequency, making it easier to tailor effective personalised marketing strategies (ibid). Fig.2 depicts the benefits of customer segmentation. Fig.3 illustrates an analysis of different BDA tools for customer segmentation.

Tool	Application	Benefits	Drawbacks
Google Analytics	Allows customer segmentation based on different factors including location, device type, age group etc. Also, it provides insights into the customer journey on the website and allows businesses to make appropriate actions to increase conversion rates (Clark,2023).	• Free of cost and easy to set up in website • Provides individual user-level details • Offers insights on conversion performance for each customer segment. (Clark,2023)	• A sample subset of the website traffic is analyzed and used to estimate the overall results. • Raw data is not accessible and offers data summary, so reports may lack valuable details, making it harder to execute sophisticated analyses. • Lack of customer support (Poddębniak and Matuszewska, 2023)
Databricks	As shown in Fig.4, Databricks can be used for customer segmentation. It is capable to process diverse data generated through a variety of e-commerce applications by applying business transformations after data cleaning. Furthermore, it will categorise customers into different segments by leveraging different types of clustering algorithms. It also provides insights into which offers and promotions are more effective for various customer segments based on their preferences (Databricks,2023a).	• Identify customer segments and which offers/promotions are suitable for them. • Accuracy, scalability, reliability and fault tolerance. • Offers flexible collaboration between different teams in the company. (Databricks,2023a)	• High costs • Not easy to setup • Require expertise and knowledge for implementation (Databricks,2023a)
Airflow and Apache Hive	Apache Hive queries can be used to gather customer details from different e-commerce applications and group the customers based on common characteristics. Airflow can be used as an orchestrator to manage the hive workflows (Hasgeek TV,2018).	• Accuracy and scalability • Fault-tolerant • Supports process rollback (Hasgeek TV,2018)	• Hive doesn't support subqueries and delete/update operations. • Hive has high latency (Simplilearn,2023b) • Require expertise and knowledge for implementation • Airflow doesn't support versioning of workflows, making development and maintenance difficult (AlexSoft,2022)

Figure 3.Analysis of BDA Tools for Customer Segmentation

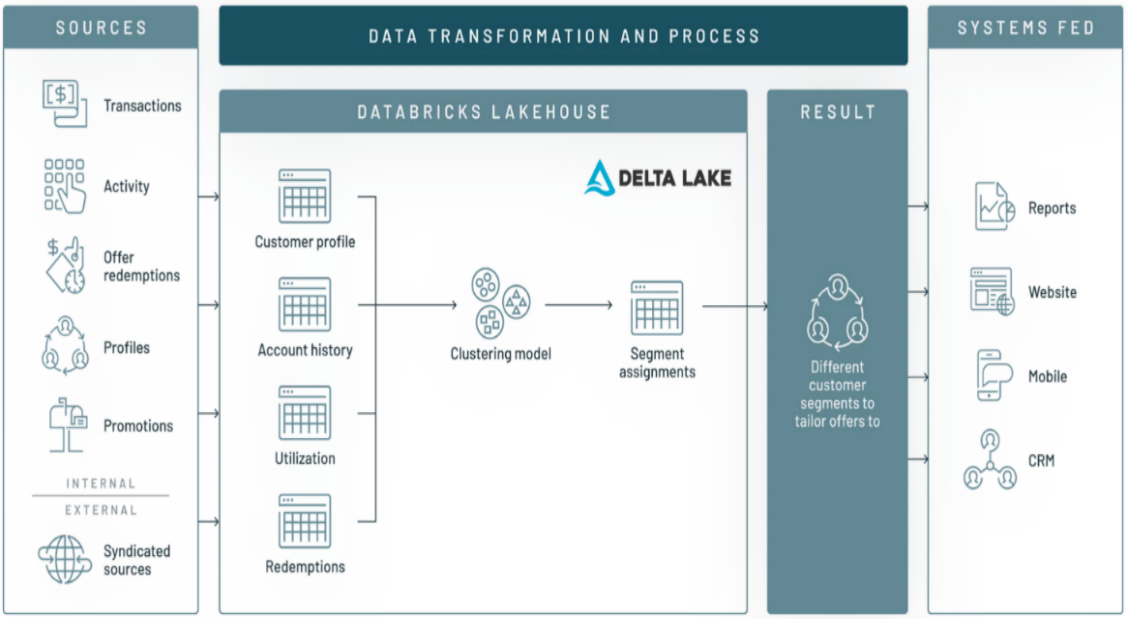

Figure 4.Sample Databricks Lakehouse Architecture for Customer Segmentation (Databricks,2023a)

A combination of Databricks and Google Analytics is recommended for Daunt Books because individual customer-level information can be captured in online transactions using Google Analytics, allowing the marketing team to further drill down for tailoring customised marketing strategies to increase conversion rates(Clark,2023). While Databricks can be used to integrate online and offline customer data for segmentation and it recommends the marketing team with suitable promotional offers for each customer segment (Databricks,2023a).

This solution will help the marketing and sales teams to overcome their challenges in tailoring marketing strategies, solving P1 from the pain points listed in Fig.2. Additionally, it partially addresses the problems P3 and P4 in Fig.2 by giving insights into the trending products among customers in different geographical locations, enabling the company to stock the right products at right outlets.

4.1.2. Inventory Management

Effective inventory management is crucial for the success of any business in the retail industry(Manning,2022). Maintaining optimum inventory levels is key to achieving efficient production and distribution of products and avoiding financial losses(Lin,2019). As illustrated in Fig.5, there are several benefits to using BDA for inventory management. Fig.6 illustrates BDA tools analysis for inventory management.

Benefits of BDA in Inventory Management	
Enhanced Operational Efficiency	• BDA helps to reduce the impact of stockouts as it can lead to lost sales and future customers, and BDA achieve this by calculating lead times and safety stock and provides insights on reorder threshold for each product through a good inventory management system. • BDA can be used to integrate SAP system and website to synchronize product stock levels, reducing the risk of overselling and allows for optimal inventory allocation across sales channels. • BDA can increase order fulfillment speed by enabling automated shipping rules, effective warehouse management, and optimization of stock movements within the warehouse to minimize picking and packing time.
Customer Satisfaction	• BDA can be used to evaluate the carrier performance and remove continuously poor performing carriers to avoid bad reputation among customers. • BDA can be used to gain insights on the factors that influence the customer shopping experience such as product descriptions, images and delivering right products without failure, enabling the company to enhance the customer experience. • BDA can be used to find the root cause for frequent returns of products such as poor product quality or damaged product and fix the issue for increased customer satisfaction.
Reduced Costs	• BDA can be help to reduce inventory management costs including warehousing, logistics, capital, storage, risk-holding, and insurance costs, by managing inventory with real-time data visibility and adopting optimization strategies.
Sales Forecasting	• BDA can be used to forecast sales based on the historical data and current market trends in different locations, allowing company to stock approximate volume of right products at right outlets to meet customer demands and generate maximum revenue during peak times. It also helps to avoid over stocking of products and product shrinkage.

Figure 5.Benefits of using BDA in Inventory Management (Choi,Wallace and Wang,2018;Manning,2022)

Tool	Application	Benefits	Drawbacks
Databricks	Databricks can be used for inventory management. Fig.7 illustrates a sample architecture for inventory management. It collects real-time and historical data from different sources including inventory, sales and returns, then book stock based on the prediction of estimated stock and sales forecast. This will help to reduce the financial impact due to overstocking and understocking of products (Databricks,2020).	• Accuracy, scalability, reliability and fault tolerance. • Ensure safety stock. • Understand the stock trends and make predictions based on the data. • Reduce financial impact due to overstocking and understocking of products. (Databricks,2020)	• High costs • Not easy to setup • Require expertise and knowledge for implementation (Databricks,2020)
Cassandra	Cassandra can be used for inventory management because it's distributed, highly responsive, and predictable scalable NoSQL database. While its low latency, ability to handle all types of data, and high read/write capacity make it ideal for providing fast and personalized customer experiences. Its multi-region replication ensures zero downtime, even in case of data centre failure. It is horizontally scalable, predictable and cost-effective. It analyses the product catalogue and inventory data in real-time and offers faster catalogue refreshes, allowing the company to limit overselling (Edwards,2021).	• Fast response times and low latency. • Capable to handle all types of high-volume data, structured and unstructured, from a variety of sources. • Scalable and cost-effective. • Easy to deploy and flexible. (Edwards,2021)	• Require expertise and knowledge • Complex querying and aggregation operations can be slow and resource-intensive. • Doesn't support ACID (Atomicity, Consistency, Isolation, Durability) properties, making it difficult to maintain data consistency and integrity in complex applications. (Simplilearn,2023a)
Tableau	Tableau can be used for inventory management because it enables users to perform data cleansing and transformation operations on a variety of data sources, including databases, spreadsheets, and cloud services. Inventory managers can easily analyse and explore complex datasets using Tableau's advanced visualisation capabilities to obtain insights into inventory levels, track inventory movements, and recognise trends or patterns. Tableau Server enables cross-organizational data exchange and collaboration on visualisations. To offer a comprehensive solution, it may be necessary to integrate it with other inventory management systems since it is primarily a data visualisation and business intelligence tool (Biswal,2023).	• Data Visualization, Exploration, Modelling, and Preparation • Collaboration and Flexibility. • Accessible through Mobiles • Supports real-time and natural language processing • User-friendly interface. • Advanced customisation options on dashboards. • Support from the community. (Biswal,2023)	• High costs • Require technical knowledge • Performance degrades with higher dataset size • Limited support on real-time processing. (Biswal,2023)
Power BI	Power BI can be used for inventory management by connecting to various data sources such as SQL Server, cleaning, transformation, and creating interactive visualizations and reports to track inventory levels, monitor inventory movements, and identify trends or patterns. The custom calculations feature in Power BI can be used to calculate reorder points, safety stock levels, and inventory turnover rates. It is accessible from different devices and integrated with other Microsoft tools like Excel and SharePoint. Additionally, Power BI can be integrated with other Microsoft tools like Excel and SharePoint to create a complete inventory management solution. With its powerful data modelling and analysis capabilities, Power BI can help organizations make data-driven decisions and optimize their inventory management processes (Biswal,2023).	• Data Visualization, Exploration, Modelling, and Preparation • Collaboration • Accessible through Mobiles • Supports fast real-time and natural language processing • Affordable price • Supports easy integration with Microsoft products. (Biswal,2023)	• Limited customisation options • Limited data connectivity • Dependence on Microsoft products (Biswal,2023)

Figure 6. Analysis of BDA Tools for Inventory Management

A combination of Databricks and Tableau is recommended for Daunt Books. Although the implementation costs for Databricks are higher than Cassandra, it offers several benefits including accuracy, scalability, reliability and fault tolerance (Databricks,2020). Furthermore, it is more suitable for inventory management since it ensures safe stock, identifies sales trends by analysing multiple data sources and makes predictions on stock estimation, thereby reducing understocking(ibid). The results from Databricks can be integrated with the website to refresh the product catalogues to restrict overselling(ibid).

Tableau can be used for data visualisations because it offers a high level of customisation options in the dashboard when compared to Power BI and can be easily integrated with cloud applications such as Databricks Lakehouse (Biswal,2023). Power BI is not recommended because it's highly dependent on Microsoft products(ibid). Therefore, a combined solution using Databricks and Tableau can overcome the problems, P3, P4 and P5 of the existing system listed in Fig.2.

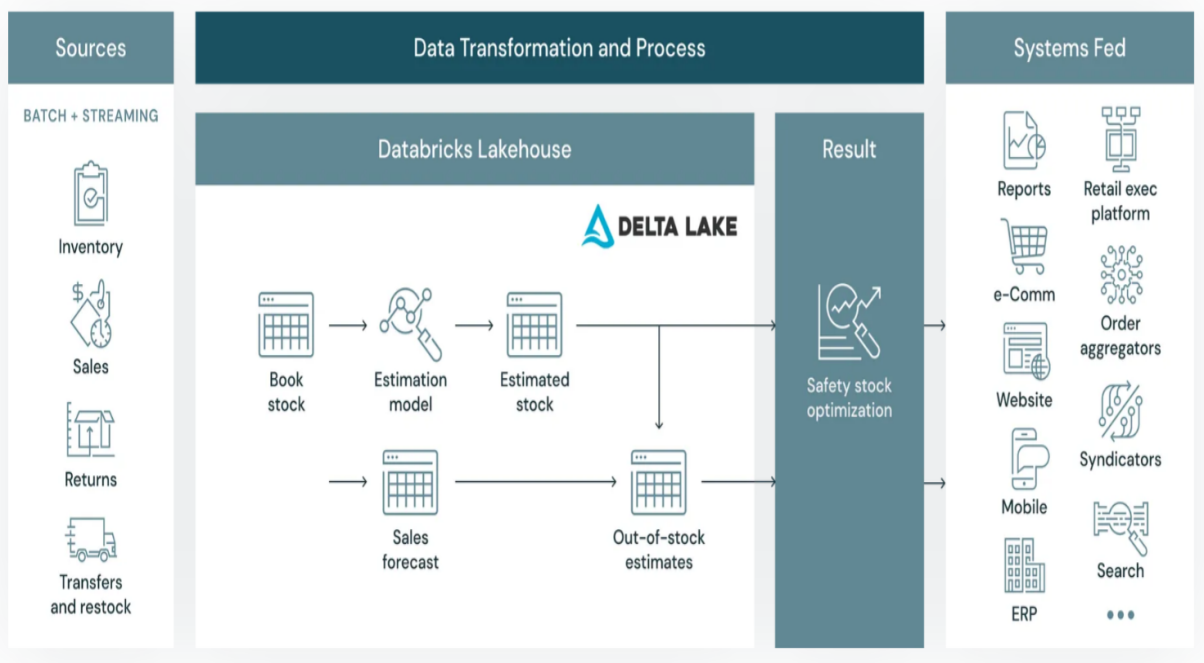

Figure 7.Sample Databricks Lakehouse Architecture for Inventory Management (Databricks,2020)

4.2. Dashboards for Evaluating Sales and Marketing Performance

Having decided on the tools for inventory management and customer segmentation, the next step is to design dashboard wireframes for tracking sales and marketing performance. Data analysts and project managers are responsible for building dashboards that assist their stakeholders in making key business decisions(Numerro ,2022). Design is frequently disregarded when selecting how these dashboards should appear and function, and analytics reports may not meet stakeholder expectations, resulting in rework and dissatisfaction for all stakeholders(ibid).

Wireframes are low-fidelity visual representations of the initial product concept that outline the page structure, layout, information architecture, and overall direction, created digitally or manually using black, white, and grey colours(Hannah,2023). Prior to analysts start building a dashboard, dashboard wireframes assist to ensure that all parties are on the same page(Numerro,2022). Fig.8 illustrates the benefits of using dashboard wireframes.

Benefits of Dashboard Wireframes	
Fast and Inexpensive	Wireframes are simple outlines of the dashboard and can be developed quickly using digital canvas or pen and paper. Therefore, they are inexpensive (Lucid,2020).
Visual Representation of Dashboard	Dashboard wireframes provide a rough idea about what needs to be included in the dashboard to meet customer expectations (Numerro,2022).
Promote Iterations	Wireframes are easily modifiable and allow designers to refine and improve the dashboard design as required (Numerro,2022).
Reduced Rework and Costs	Early feedback on the dashboard designs helps to reduce rework and costs (Numerro,2022).
Encourage Collaboration and Revision of Dashboard Design	Dashboard wireframes assist stakeholders to convey their ideas and opinions in the dashboard design before analysts start building a dashboard. This will reduce rework, save time and effort, and revise the dashboard design that aligns with the business objectives and scope (Numerro,2022)

Figure 8.Benefits of Dashboard Wireframes

4.2.1. Key Performance Indicators

Before designing wireframes, it is important to identify the Key Performance Indicators(KPIs) that need to be included in the dashboard. KPIs are metrics that measure activity with significant business impact, enabling everyone to stay aligned on the metrics contributing to company growth and ensure the team is achieving the organisational goals(Salesforce,2022). Fig.9 and Fig.10 illustrate some examples of Sales and Marketing KPIs respectively, that enable company CEO to make data-driven decisions.

Sales KPI	Description	Calculation	Benefit
Total Sales Revenue	The total amount of sales revenue generated during the given period of time – daily, monthly or yearly. It helps to understand sales trends.	Sales Revenue= *SUM(Quantity*Unit Price)*	Allows to understand the overall financial performance of the company. Also, it helps to track progress towards revenue goals, identify trends in sales performance, and make data-driven decisions about the company's sales strategy.
Customer Lifetime Value (CLV)	The sum of all purchases that a customer makes throughout their relationship with the company, including upsells, cross-sells, and renewals. (Salesforce,2022)	CLV= *(Avg. purchase value per year) x (Avg. no. of purchases per year for each customer) x (Avg. customer lifespan in years)*	CLV measures how well the company establishes trusting, value-focused, and loyal customer relationships that result in renewals, upsells, and cross-sells, leading to predictable income. (Salesforce,2022)
Customer Retention	The percentage of customers who continue to buy books from the company during a year. (Salesforce,2022)	Customer Retention= *(Total no. of customers at the end of the year – Net new customers acquired during the year) / (No. of customers at the start of the year) x 100*	Focusing on customer retention and upselling/cross-selling can lead to predictable revenue and maximize ROI, while falling retention may require revisiting customer engagement strategies to ensure that company representatives are giving priorities to loyal customers. (Salesforce,2022)
Customer Churn Rate	The percentage of customers who stopped purchasing books from the company during a month. (Salesforce,no date)	Customer Churn Rate= *(No. of customers at the beginning of the month – No. of customers at the end of the month) / (No. of customers at the beginning of the month)*	To assess the success of your marketing initiatives and customer satisfaction, it is critical to understand customer churn because maintaining existing clients is simpler and less expensive than finding new ones. (Salesforce,no date)
Gross Profit Margin	The percentage of revenue that remains after deducting the cost of products sold. (Indeed,2023)	Gross Profit Margin = *(Revenue – Cost of products sold) / Revenue*	Helps to measure the company's profit margin. A high-profit margin indicates that the CEO is making the right business decisions, helping the company to strengthen through investing or saving its profits. While a lower profit margin indicates that the CEO needs to formulate a new strategy to increase overall revenue or reduce costs. (Indeed,2023)

Figure 9.Examples of Sales KPIs

Marketing KPI	Description	Calculation	Benefit
Returns On Investment (ROI)	It is the measure of the profitability of an investment over a given period of time (monthly, quarterly or annually) (Gillis,2021)	ROI= *Net income / Cost of investment x 100.*	Evaluates the effectiveness of the company's investments and make informed decisions about allocating resources. By calculating ROI, the CEO can determine whether the company's investments are generating sufficient returns and adjust the strategies accordingly. A positive ROI indicates that the investment is profitable, while a negative ROI indicates that the investment is generating losses (Gillis,2021).
Customer Acquisition Costs (CAC)	Total sales and marketing costs spent to earn a new customer over a given period of time (monthly, quarterly or annually) (Bernazzani,2023)	CAC= *(Cost of Sales + Cost of Marketing)/New Customers Acquired*	Understanding CAC, the CEO can determine the profitability of their business model and adjust their marketing and sales strategies to ensure that they are acquiring customers cost-effectively. Additionally, tracking the CAC over time can help the CEO to identify trends and make informed decisions about future investments in sales and marketing activities. (Bernazzani,2023)
Net Promoter Score (NPS)	It is the measure of customer loyalty and satisfaction with a company's product or service (Bernazzani,2022)	NPS=*(Number of Promoter Scores/Total Number of Respondents) - (Number of Detractor Scores/Total Number of Respondents)*	Provides insight into how likely customers are to recommend the company to others, which can significantly impact the company's growth and success. A high NPS score indicates that customers are satisfied and loyal resulting in higher brand value, while a low score may suggest that the company needs to make improvements to its product or service to promote brand value (Bernazzani,2022).
Marketing ROI	It is the measure of the effectiveness and profitability of the company's marketing strategies over a given period of time (monthly, quarterly or annually) (Beattie,2022)	Marketing ROI=*(Gross Marketing Revenue – Marketing Investment)/ Marketing Investment*	Evaluating the effectiveness of the company's marketing strategies and the return on the investments spent on marketing activities. A higher Marketing ROI indicates that the marketing campaigns are generating more revenue for the company than the money spent on them, which is a positive sign for the company's profitability. By tracking the Marketing ROI, the CEO can make data-driven decisions on where to allocate the marketing budget and how to optimize the marketing strategies to improve the ROI (Beattie,2022)
Conversion Rate	It is the percentage of unique customers who achieved a desired action or goal set by the business. It can be purchasing a book or an item from gift collections (Johnson, no date).	Conversion Rate = *Total number of conversions / Total number of unique visitors* 100*	Evaluates the effectiveness of the company's marketing and sales efforts in turning potential customers into actual customers. By tracking conversion rates, the CEO can identify areas of improvement in the sales and marketing process and make data-driven decisions to optimize business performance (Johnson, no date).

Figure 10.Examples of Marketing KPIs

4.2.2. Sales and Marketing Performance Dashboard Wireframes

Based on the relevant KPIs for marketing and sales that are relevant to CEO that were discussed in the previous section, Fig.11 and Fig. 12 illustrates the sample dashboard wireframes for sales and marketing performance respectively.

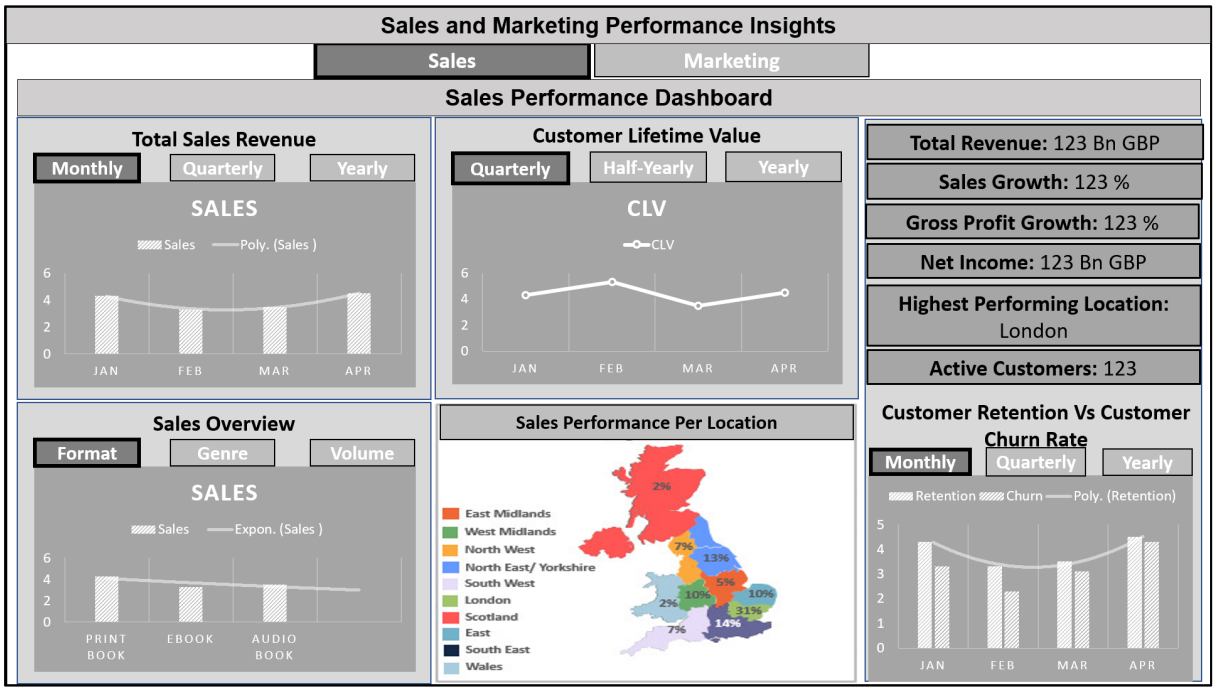

Figure 11.Sales Performance Dashboard Wireframe Example(Calzon,2022)

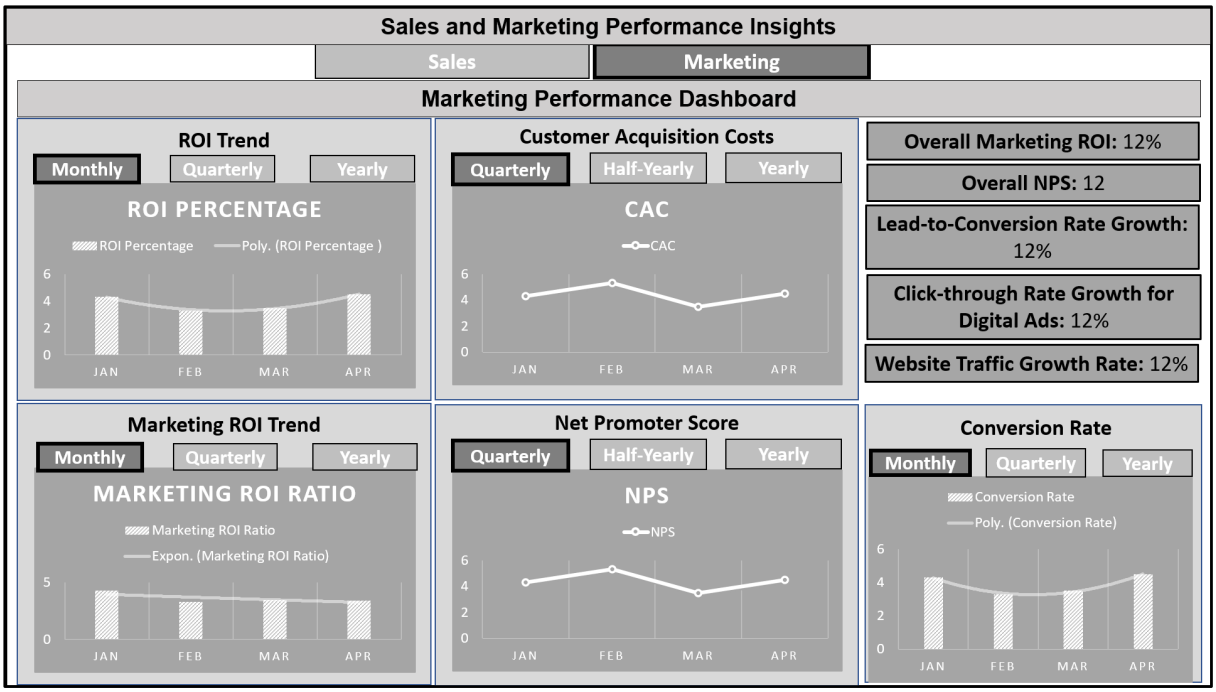

Figure 12.Marketing Performance Dashboard Wireframe Example(Calzon,2022)

5. Action Plan for Daunt Books

Having identified the tools and metrics for enhancing and monitoring sales and marketing performance, this section provides a high-level action plan for Daunt Books for implementing the proposed system as illustrated in Fig.13.

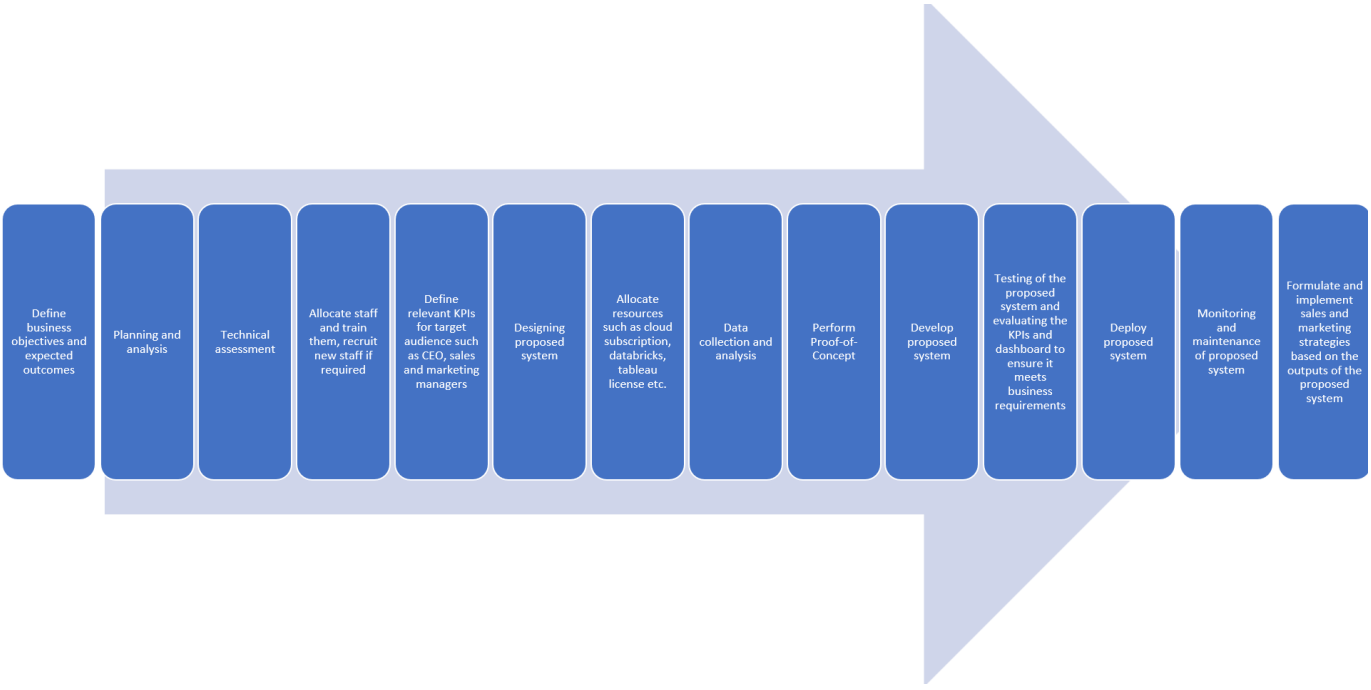

Figure 13.High-level Action Plan for Proposed System Implementation

The complete implementation of the action plan might take from 1 year to 2 years since it includes technical complexity and need to undergo several processes.

5.1.1. Risks Involved and Mitigation Actions

Fig.14 depicts the risks involved in the implementation of the proposed system and suggests actions to mitigate them.

Risk Category	Risk Description	Risk Level	Mitigation Action
Adoption	The adoption of BDA solutions might not add any value to the business if it is not implemented correctly without a clear understanding of the business requirements, expected outcomes and risks involved in it (Elia *et al.*,2020).	Moderate	A thorough analysis should be conducted before adopting the BDA solution to any business use case to ensure that it aligns with the business requirements and can produce expected outcomes(Elia *et al.*,2020). It should be carefully designed and implemented to avoid any undesired result(ibid). Furthermore, it should be continuously monitored and evaluated after adoption to check whether it is adding value to the business and to identify improvement areas(ibid).
Privacy and Security	Google Analytics may collect Personally Identifiable Information (PII) from customers who visit the company website, such as their email addresses, and demographic information, which could be a potential breach of privacy and security(Duncan,2023).	High	The company should adopt security measures including configuring Google Analytics to mask all user IP addresses, leveraging data retention policies to delete user data after the retention period, and using Google Tag Manager to have more control over the data being collected. Additionally, the company can update its privacy policy to inform customers of the data being collected and how it is being used, obtain consent from them as well as offer them disable option from being tracked by Google Analytics. The company must conduct regular audits to ensure that the implementation of Google Analytics is in compliance with privacy regulations such as GDPR(General Data Protection Regulation) (Duncan,2023).
Implementation	Effective implementation of Databricks and Tableau requires expertise and knowledge. Furthermore, it might lead to implementation delays or extra costs due to technical complexities (Databricks,2020;Biswal,2023)	Moderate	The company should train the staff on the new tools that are under consideration and employ technical experts if required. Also, the company can seek technical support from Databricks in the implementation and maintenance of the proposed system. Proper planning, training, analysis and testing before implementation can help to mitigate the implementation risks. (Databricks,2020)
Technical	The performance of Tableau may degrade with huge volume of data and has limited support for real-time streaming(Biswal,2023)	Moderate	The performance of Tableau can be monitored using Tableau Performance Recording which enables the technical team to identify the bottleneck and areas of improvement in the workbook (Tableau,no date). Databricks can process the data in real-time using Streaming and thereby handle the shortcomings of Tableau in real-time processing (Databricks,2023b). Furthermore, Databricks can process data in the incremental method by reducing the workload in Tableau during visualisations (Databricks,2023a).
Implementation	The data captured through the website can be incomplete resulting in data quality issues. This may lead to incorrect predictions or outcomes from the proposed system.	Low	Databricks ensures data quality by a transactional storage layer, Delta that includes data quality features such as schema enforcement and data validation rules etc. (Databricks,2023c)

Figure 14.Risks Involved and Mitigation Actions

6. Discussion

As the company grows it has to handle huge volume of customer data for analytics to make data-driven decisions. Currently the company is following DevOps methodology, it can move towards DevSecOps culture that will integrate the security into each iteration of SDLC, ensuring the system is following data privacy and security regulations (Sugandhi,2022). Furthermore, the proposed system can adopt more sophisticated applications including sales forecasting, fraud detection, predictive analysis for product recommendation and predictive search using BDA and ML techniques in future (Alrumiah and Hadwan,2021). This will enhance the customer experience further and generate more revenue and conversion rates.

7. Conclusion

The requirement of Daunt Book's CEO was to propose a strategy that integrate the channels, enhance marketing strategies and optimise sales using BDA technology. As the initial step, this study analysed the existing system, organisation structure, work culture and identified improvement areas. Customer segmentation and inventory management were the two main focus areas of this study for technical modernisation. A combination of Databricks and Google Analytics were proposed for customer segmentation. While Databricks and Tableau were proposed for inventory management. This report also covered the relevant KPIs for CEO that helps to monitor the sales and marketing performance. This report highlighted the benefits, challenges, and mitigation actions of the suggested approach. Additionally, the report outlined a high-level action plan for Daunt Books and discussed future improvements for the proposed system.

8. References

Alexander, G. (2023) *UK books and E-books market report - market size & forecast*, *Mintel Store*. Mintel. Available at: https://reports.mintel.com/display/1100541/?fromSearch=%3Ffreetext%3Dbooks%26resultPosition%3D1 (Accessed: April 10, 2023).

AlexSoft (2022) *The good and the bad of Apache Airflow Pipeline Orchestration*, *AltexSoft*. AltexSoft. Available at: https://www.altexsoft.com/blog/apache-airflow-pros-cons/ (Accessed: April 22, 2023).

Alrumiah, S.S. and Hadwan, M. (2021) "Implementing big data analytics in e-commerce: Vendor and Customer View," *IEEE Access*, 9, pp. 37281–37286. Available at: https://doi.org/10.1109/access.2021.3063615.

Beattie, A. (2022) *How to calculate the return on investment (ROI) of a marketing campaign*, *Investopedia*. Investopedia. Available at: https://www.investopedia.com/articles/personal-finance/053015/how-calculate-roi-marketing-campaign.asp (Accessed: April 26, 2023).

Behl, A. *et al.* (2019) "A conceptual framework for the adoption of big data analytics by e-commerce startups: A case-based approach," *Information Systems and e-Business Management*, 17(2-4), pp. 285–318. Available at: https://doi.org/10.1007/s10257-019-00452-5.

Bernazzani, S. (2023) *Customer acquisition cost: How to calculate CAC [+benchmarks & formulas to know]*, *HubSpot Blog*. HubSpot. Available at: https://blog.hubspot.com/service/what-does-cac-stand-for#:~:text=You%20can%20calculate%20customer%20acquisition,Number%20of%20New%20Customers%20Acquired. (Accessed: April 26, 2023).

Bernazzani, S. (2022) *How to calculate net promoter score [formula]*, *HubSpot Blog*. HubSpot. Available at: https://blog.hubspot.com/service/how-to-calculate-nps (Accessed: April 26, 2023).

Biswal, A. (2023) *Microsoft power bi vs tableau: Here are the difference you should know: Simplilearn*, *Simplilearn.com*. Simplilearn. Available at: https://www.simplilearn.com/tutorials/power-bi-tutorial/power-bi-vs-tableau (Accessed: April 22, 2023).

Bython (2023) *Understanding the benefits of customer segmentation*, *Bython*. Available at: https://bython.com/benefits-of-customer-segmentation/ (Accessed: April 17, 2023).

Calzon, B. (2022) *CEO Dashboard Examples & Reports with leadership metrics*, *datapine*. Available at: https://www.datapine.com/blog/ceo-dashboard-report-examples-and-templates/ (Accessed: April 26, 2023).

Chiang, L.-L. and Yang, C.-S. (2018) "Does country-of-origin brand personality generate retail customer lifetime value? A big data analytics approach," *Technological Forecasting and Social Change*, 130, pp. 177–187. Available at: https://doi.org/10.1016/j.techfore.2017.06.034.

Choi, T.-M., Wallace, S.W. and Wang, Y. (2018) "Big Data Analytics in Operations Management," *Production and Operations Management*, 27(10), pp. 1868–1883. Available at: https://doi.org/10.1111/poms.12838.

Clark, H. (2023) *10 best customer segmentation tools in 2023*, *The CX Lead*. Available at: https://thecxlead.com/tools/best-customer-segmentation-tools/ (Accessed: April 16, 2023).

Databricks (2020) *Solution accelerator - how to build: Safety Stock*, *Databricks*. Available at: https://www.databricks.com/solutions/accelerators/safety-stock (Accessed: April 19, 2023).

Databricks (2023a) *Solution accelerator - how to build: Customer segmentation for personalization*, *Databricks*. Available at: https://www.databricks.com/solutions/accelerators/customer-segmentation (Accessed: April 16, 2023).

Databricks (2023b) *Streaming on Databricks*, *Streaming on Databricks | Databricks on AWS*. Available at: https://docs.databricks.com/structured-streaming/index.html (Accessed: April 23, 2023).

Databricks (2023c) *What is Delta Lake?*, *What is Delta Lake? | Databricks on AWS*. Available at: https://docs.databricks.com/delta/index.html (Accessed: April 23, 2023).

Duncan, Z. (2023) *GA4 problems - 4 challenges with google analytics 4*, *Digital Marketing and Analytics | Root and Branch*. https://www.rootandbranchgroup.com/wp-content/uploads/2018/06/Black_GLeaf_2-300x248.png. Available at: https://www.rootandbranchgroup.com/ga4-problems/ (Accessed: April 23, 2023).

Edwards, R. (2021) *What are people using Cassandra for anyway?*, *DataStax*. DataStax. Available at: https://www.datastax.com/blog/exploring-common-apache-cassandra-use-cases (Accessed: April 22, 2023).

Elia, G. *et al.* (2020) "A multi-dimension framework for value creation through Big Data," *Industrial Marketing Management*, 90, pp. 617–632. Available at: https://doi.org/10.1016/j.indmarman.2020.03.015.

Errera, R. (2022) *Eye-popping book and reading statistics [2022]*, *Toner Buzz*. Available at: https://www.tonerbuzz.com/blog/book-and-reading-statistics/ (Accessed: April 15, 2023).

Gillis, A.S. (2021) *What is roi? how to calculate return on investment*, *CIO*. TechTarget. Available at: https://www.techtarget.com/searchcio/definition/ROI#:~:text=The%20most%20common%20is%20net,Cost%20of%20investment%20x%20100. (Accessed: April 26, 2023).

Hannah, J. (2023) *Wireframe vs prototype vs mockup: We explain the difference*, *CareerFoundry*. Available at: https://careerfoundry.com/en/blog/ux-

design/difference-between-wireframes-prototypes-mockups/ (Accessed: April 24, 2023).

Hasgeek TV (2018) *Segmenting 500 million users using Airflow + Hive - Soumya*. YouTube. Available at: https://www.youtube.com/watch?v=ibWAgVbDXow (Accessed: April 16, 2023).

Indeed (2023) *8 possible CEO KPIs to measure CEO and Business Performance*, *Indeed Career Guide*. Indeed. Available at: https://www.indeed.com/career-advice/career-development/ceo-kpi (Accessed: April 24, 2023).

Jocelyn, V. and Biagi, L. (2022) *Book market in the United Kingdom (UK)*, *Statista*. Available at: https://www.statista.com/study/21970/book-market-in-the-united-kingdom-statista-dossier/ (Accessed: April 12, 2023).

Johnson, B. (no date) *How to calculate conversion rate - the conversion rate formula*, *How To Calculate Conversion Rate - The Conversion Rate Formula*. Available at: https://blog.useproof.com/calculate-conversion-rate (Accessed: April 26, 2023).

Lin, R. (2019) "2019 International Conference on Management, Finance and Social Sciences Research (MFSSR 2019)," *The Importance of Successful Inventory Management to Enterprises-A Case Study of Wal-Mart*, pp. 755–758. Available at: https://doi.org/https://webofproceedings.org/proceedings_series/ECOM/MFSSR%202019/MFSSR19154.pdf.

Lucid (2020) *Wireframes vs mockups: Determining the right level of Fidelity for Your Project*, *Wireframes vs Mockups Explained | Lucidchart Blog*. Available at: https://www.lucidchart.com/blog/wireframes-vs-mockups (Accessed: April 24, 2023).

Manning, A. (2022) *How big data can improve inventory management*, *The BigCommerce Blog*. Available at: https://www.bigcommerce.co.uk/blog/data-inventory-management/ (Accessed: April 17, 2023).

Numerro (2022) *The Complete Guide to designing a BI dashboard mockup*, *Power BI Dashboard Design Toolkit from Numerro*. Available at: https://www.numerro.io/guides/dashboard-mockup-guide (Accessed: April 23, 2023).

Poddębniak, M. and Matuszewska, K. (2023) *6 key google analytics limitations*, *Piwik PRO*. Available at: https://piwik.pro/blog/google-analytics-drawbacks/#what-are-the-limitations-of-google-analytics-4? (Accessed: April 22, 2023).

Salesforce (2022) *9 sales kpis every sales team should be tracking*, *Salesforce*. Available at: https://www.salesforce.com/resources/articles/sales-kpis/ (Accessed: April 24, 2023).

Salesforce (no date) *How to calculate customer churn rate and revenue churn rate*, *Salesforce*. Available at: https://www.salesforce.com/resources/articles/how-calculate-customer-churn-and-revenue-churn/#:~:text=How%20do%20you%20calculate%20customer%20churn%20rate%3F,additional%20revenue%20from%20existing%20customers. (Accessed: April 25, 2023).

Sayyad, S. *et al.* (2019) "Digital Marketing Framework Strategies through big data," *Lecture Notes on Data Engineering and Communications Technologies*, pp. 1065–1073. Available at: https://doi.org/10.1007/978-3-030-24643-3_127.

Simplilearn (2023a) *Cassandra vs. mongodb: The top differences you should know*, *Simplilearn.com*. Simplilearn. Available at: https://www.simplilearn.com/cassandra-vs-mongodb-article (Accessed: April 22, 2023).

Simplilearn (2023b) *What is Hive?: Introduction to hive in hadoop: Simplilearn*, *Simplilearn.com*. Simplilearn. Available at: https://www.simplilearn.com/what-is-hive-article#limitations_of_hive (Accessed: April 22, 2023).

Sugandhi, A. (2022) *DevOps VS DevSecOps: Top differences*, *KnowledgeHut*. Knowledgehut. Available at: https://www.knowledgehut.com/blog/devops/devops-vs-devsecops (Accessed: April 27, 2023).

Tableau (no date) *Record and analyze workbook performance*, *Online Help*. Available at: https://help.tableau.com/current/pro/desktop/en-us/perf_record_create_desktop.htm (Accessed: April 23, 2023).

B. PART-2(B): Data Architecture Proposal for Joules

1. Introduction

The clothing industry relies on data analytics for several reasons including inventory management, trend and sales forecasting, customer segmentation and targeting, personalisation and data-driven decisions(Hickins,2023). The data analytics process is computationally intensive and requires high-quality data to provide accurate insights to the target audience(Rehan,2022). Data analytics help clothing brands to offer curated content, style advice and personalised shopping experience for their customers based on their lifestyles and values(SAP, no date). As part of Industry 4.0, more clothing companies are adopting digital technologies for data analytics to gain a competitive edge in the market(Ahmad *et al.*,2020).

2. Organisation Insights

Joules is a medium-sized UK clothing company that operates online and offline with more than 100 stores across the UK. It offers a wide variety of clothes and accessories suitable for men, women and kids as well as homewares, focusing on sustainability and ethical production practices. Joules is known for its jackets and patterned Wellington boots, however, it's experiencing a significant decline in sales due to rising living costs and climate change in the UK(Jolly,2022). Furthermore, the company is not profitable due to higher freight charges and wage costs(Provan and Wallace,2022). Therefore, the company's CEO hired an external consultant to conduct a technical assessment and recommend the best solution using big data and analytics (BDA) technology for overcoming the challenges.

Through this assessment, the CEO attempts to seek the answer to the following questions:

- How can BDA technology aid to improve sales and profitability?
- How can BDA technology be leveraged to optimize the supply chain and reduce operational costs?
- How can BDA technology help the company to reduce wastage and promote the recycling of products?
- Provide a high-level design(HLD) for the proposed system.

- What are the risks involved in the implementation of the proposed system and mitigation steps?

The answers to the above questions will be covered in the upcoming sections.

2.1. Organisation Structure

The company has dedicated teams for IT, Sales, Marketing, Customer Support, Operations and Supply Chain. The IT team is following the DevOps methodology for delivering software products. Sales and Marketing teams have a set of targets that need to be achieved at regular intervals of time. The company manages deliveries from all sourcing locations in Asia to the UK manufacturing unit, including managing all customs and compliance requirements through a third-party company, SEKO Logistics(SEKO,no date). Joules have partnerships with Reskinned and Oxfam for recycling the products to reduce wastage(Joules,no date).

3. Market Research

Fig.15 depicts the key highlights of the market research. Based on the research, Joules has business opportunities despite challenges since it is focusing on sustainability and recycling practices. Therefore, data analytics will help Joules to gain more insights into business data, allowing it to achieve a competitive edge in the market and increase the total sales revenue.

Market Research Highlights	
Top 3 Competitors	Boden, Crew Clothing, FatFace (SimilarWeb,2023)
Customer Preferences	• Cheaper price. • Prefer warm clothes over cold ones in winter due to increasing energy charges. • Online shopping has increased after the pandemic. • Tends to purchase from brands that offer better loyalty and reward schemes. • Increased demand for recycled, sustainable and second-hand products. (Ceron,2023) • UK summer wave has reduced the demand for Wellington boots and jackets (Jolly,2022).
Latest Trends	• Focus more on Sustainability and Recycling (Ceron,2023) • Decline in overall clothing purchases due to rising costs(ibid) • AI is used for providing style advice to customers based on their lifestyles. (Zavialova,2023)
Sales Driving Factors	• Cost of Living Crisis • Clothing Inflation • Climate Change • Unemployment (Ceron,2023)
Opportunities	Older women tend to purchase more than young women, despite the inflation and rising costs. However, most brands mainly focus on youngsters making it difficult for older customers to find suitable clothes. (Ceron,2023)

Figure 15.Market Research Highlights

4. Business Objectives and Challenges of Joules

It is important to understand the business objectives and challenges of the company for proposing an effective solution. Fig.16 lists the business requirements that need to be addressed using BDA technology.

Problem No.	Affected Team	Pain Point	Identified Root Cause
P1	Sales & Marketing	Although customers prefer lower-priced products in the current cost of living crisis, the company is unable to reduce the product cost, resulting in a significant fall in sales (Olabiyi,2022).	High freight charges and wage costs resulted in increased product price(Provan and Wallace,2022).
P2	Operations	Difficulty in managing the stocks as per the customer demand and latest fashion trends(Olabiyi,2022).	Improper inventory management, sales and fashion trend forecasting.
P3	Design	With the increasing focus on sustainability and ethical practices, seeking innovative approaches to reduce wastage and promote recycling of products with new designs(Joules, no date).	Designers don't have proper insights into rapidly changing customer preferences and are unable to forecast the trends, resulting in deadstock.
P4	Supply Chain	Difficulty in reducing freight charges and labour costs to increase the company's profit margin (Provan and Wallace,2022).	Warehousing and shipping operations are not optimised. Also, the staff is not allocated properly as per the warehouse and delivery requirements.

Figure 16.Business Objectives and Challenges of Joules

Having understood the business requirements, the next section recommends the proposed system using BDA technology.

5. Proposed System using Big Data and Analytics Technology

BDA technology is suitable for addressing the challenges faced in the clothing industry. Fig.17 illustrates some benefits of using BDA technology in this industry.

Benefits of BDA in the Clothing Industry	
Customer Satisfaction and Retention	Helps to retain online and offline customers by enabling the company to fulfil their changing preferences such as fashion trends, price, product material etc. Also, BDA can enhance customer satisfaction through customer segmentation and thereby offer personalised experiences such as product recommendations, promotions etc. (Bython,2023;Certilogo,2022)
Good Customer Relationships	The company can identify their loyal and premium customers from the huge database that contains information regarding online and offline customers and also allows them to keep a good relationship with them by offering special loyalty services, and discount coupons for recycling products(Bython,2023;Joules,no date).
Enhanced Competitiveness	A clear understanding of the target audience preferences gives the company a competitive edge over its competitors such as innovative and sustainable designs, competitive prices etc (Certilogo,2022).
Price Optimisation	The company can optimise product prices to an extent by understanding the budget of the target audience for purchase, resulting in increased sales and conversion rates. It is crucial as customers are going through a cost of living crisis and climate change as well as being conscious about sustainability. (Bython,2023;Certilogo,2022;Jolly,2022)
Enhanced Distributed Sales Channel	Helps to identify customer shopping behaviour and geographical preference for distributing products to stores accordingly. Collecting data on the product movement within each store using RFID to gain more insights into which are the right products that need to be stocked in each store. (Bython,2023;Olabiyi,2022)
Data-Driven Decisions for Innovative Designs	Several factors influence product designs such as customer preference and emotions, seasons, fashion trends, occasions, themes etc. BDA can assist designers to make data-driven design decisions which align with customer preferences and market trends based on real-time and historical data. Also, it will help designers to reduce the wastage of raw materials and promote sustainable and ethical practices as well as recycling products.(Certilogo,2022; Kiron,2021)
Personalised Marketing Strategies	Helps to tailor marketing strategies that could attract potential customers for repeated purchases. (Bython,2023)
Increased Revenue and Profit Margin	Enables to increase sales revenue by marketing the specific target audience preferences such as designer products, reducing operational costs and wastage(Certilogo,2022;Olabiyi,2022)
Demand Forecasting	Helps to forecast demands based on historical data and current market trends in different locations, allowing the company to stock the estimated volume of the right products at the right stores to meet customer demands and generate maximum revenue during peak times. It also helps to avoid product deadstock and shrinkage. Also, can identify new opportunities by uncovering hidden trends. (Certilogo,2022; Olabiyi,2022)
Product Recommendation	BDA can offer product recommendations based on past customer shopping behaviour (Lin et. al,2022).
Enhanced Operational Efficiency and Reduced Costs and Wastage	Helps to optimise operations and supply chain management by identifying improvement areas and adopting optimisation strategies including transportation routes, reducing freight charges and labour costs(Lamba and Singh,2017). Also, helps to stock the estimated amount of the right products through proper forecasting (Certilogo,2022;Olabiyi,2022).

Figure 17.Benefits of BDA Technology in the Clothing Industry

5.1. High-level Design(HLD) for the Proposed System

It is crucial to have a clear understanding of the data sources, data format, storage, and usage to propose an effective solution for the company(Olabiyi,2022). These are covered in detail in the following sections.

5.1.1. Data Sources

Joules deals with a wide variety of batch and real-time data from diverse applications. Fig.18 illustrates the key data sources, purpose, data collection frequency and format.

Data Source	Purpose	Data Collection Frequency	Data Format
SAP System	SAP modules are used for different business operations such as Customer Relationship Management(CRM), Enterprise Resource Planning(ERP), Point Of Sales(POS) and Supply Chain Management(SCM).	Real-time and Batch	Structured
Signals from RFID chips	Leverages RFID chips on the products to gather real-time data related to product movement(Olabiyi,2022). For example, the frequency of inward and outward product movement from trial rooms or the duration of time that each product takes to reach POS(ibid).	Real-time	Semi-structured
Social Media Platforms	Monitors social media platforms to understand the changing customer requirements and preferences.	Real-time	Unstructured
Website Analytics	Joules leverages different tools for website analytics (BuiltWith, no date). For example, Google Analytics is used to gain insights into individual customer-level information that will help the company to tailor personalised sales and marketing strategies(Clark,2023).	Real-time and batch	Semi-structured, Structured
Logs	Stores various types of logs including customer call logs and online chat logs for resolving customer problems(Hickins,2023).	Real-time	Semi-structured
MS SQL Server	Stores transactional data and aggregates data from different sources such as ERP and POS for preparing reports.	Real-time and batch	Structured

Figure 18.Key Data Sources for Joules

Having identified the data sources, the next step is to select a suitable data storage strategy.

Features	Data Warehouse	Data Lake	Data Lakehouse
Storage Data Type	Works well with structured data	Works well with semi-structured and unstructured data	Can handle structured, semi-structured, and unstructured data
Purpose	Best for data analytics and business intelligence (BI) use-cases	Suitable for machine learning (ML) and artificial intelligence (AI) workloads	Suitable for both data analytics and ML workloads
Cost	Storage is costly and time-consuming	Storage is cost-effective, fast, and flexible	Storage is cost-effective, fast, and flexible
ACID *(Atomicity, Consistency, Isolation and Durability)* **Compliance**	Records data in an ACID-compliant manner to ensure the highest levels of integrity	Non-ACID compliance and updates and deletes are complex operations	ACID-compliant to ensure consistency as multiple parties concurrently read or write data
Performance	High	Low	High
Data Formats	Closed	Open data such as Parquet, ORC, AVRO, etc.	Open and standardized formats such as Parquet, ORC, AVRO, etc.
Data governance	Simple and well-defined data governance	Poor governance capabilities	Complex but well-defined data governance
Scalability	Limited scalability and it becomes expensive	High Scalability	High Scalability
Use cases	Business Intelligence(BI) and Reporting	Data Analytics, Data Science, ML, and AI	All use cases of data warehouse and data lakes

Figure 19.Comparison between Data Warehouse, Data Lake and Data Lakehouse (Kutay,2023;Logunova,2022;Wickramasinghe,2023)

5.1.2. Data Storage Architecture

The most commonly used data storage architectures are data warehouse, data lake and data lakehouse(Longunova,2022). Fig.19 illustrates the comparison between these three architectures.

Data Lakehouse is recommended for Joules because it enables the company to bring together all the relevant data from different sources into a single data source, making analytics easier since data is stored in silos such as SAP ERP, SAP CRM etc. in the existing system(Kava and Gong,2021). It allows concurrent read/write of transactional data using Delta Engine and ensures data is ACID compliant which is crucial in eCommerce (Balasubramaniam,2022;Kutay,2023). Furthermore, it is cost-effective because the company have to manage only a single data source and reduce data redundancy(Logunova,2022). Additionally, it enforces schema and data integrity, enabling easy implementation of data security and governance mechanisms using Unity Catalog(Balasubramaniam,2022;Kutay,2023). It offers easy integration with other applications including BI and ML tools, enabling the company to leverage data visualisations using BI, and sales and demand forecasting using ML(Logunova,2022).

5.1.3. Data Lakehouse Layers

Data Lakehouse follows a layered architecture and the functionalities of each layer are depicted in the Fig.20.

Layer	Functionality	Tools Examples
Ingestion	Collects both batch and streaming data in different formats from diverse sources and delivers it to the storage layer.	**Data Streaming:** *Apache Kafka, AWS Kinesis* **Batch Processing:** *Apache Spark, AWS Data Migration Service*
Storage	Allows to store raw data as objects in different low-cost storage mediums and allows client applications to read data from the object store using open file formats such as Parquet.	*AWS S3,HDFS*
Metadata	Offers a unified catalogue that provides metadata for all objects stored in a data lake and allows users to apply management features, including ACID transactions, cache, indexing, and data extraction. It also allows the application of schema architectures such as star and snowflake schemas, implements schema management, and provides data governance and auditing functionality directly on the data lake, enhancing the overall data pipeline quality.	*Delta Lake, Unity Catalog*
API	Provides various APIs that enable end users to process their tasks faster and also offer access and access methods to the required data for advanced analytics.	*DataFrame API, SQL API*
Consumption	Offers various tools and applications in the consumption layer for different analytics tasks including business dashboards or ML analytics.	*Power BI, Tableau and ML applications*

Figure 20.Data Lakehouse Layers(Alexsoft,2021)

5.1.4. Data Lakehouse Architecture

Databricks Data Lakehouse is recommended for Joules because it can be implemented on any cloud platform such as AWS, Azure or GCP(Databricks,no date). Databricks offer customer support and have good documentation that assists the technical team during and after the implementation(ibid). It offers high performance while executing complex queries in real-time by optimising Apache Spark APIs and leveraging Photon, the next-generation lakehouse engine compatible with Spark APIs(ibid). Databricks offers orchestration through workflows and MLFlow for MLOps(ibid). Fig.21 illustrates the HLD for Databricks Data Lakehouse.

Although AWS offers a data lakehouse solution, it is not recommended because the implementation of data lakehouse is not easy as Databricks and doesn't offer MLFlow, Unity Catalog or Photon Engine integrations that offer higher performance(Kava and Gong,2021;Kossendey,2023).

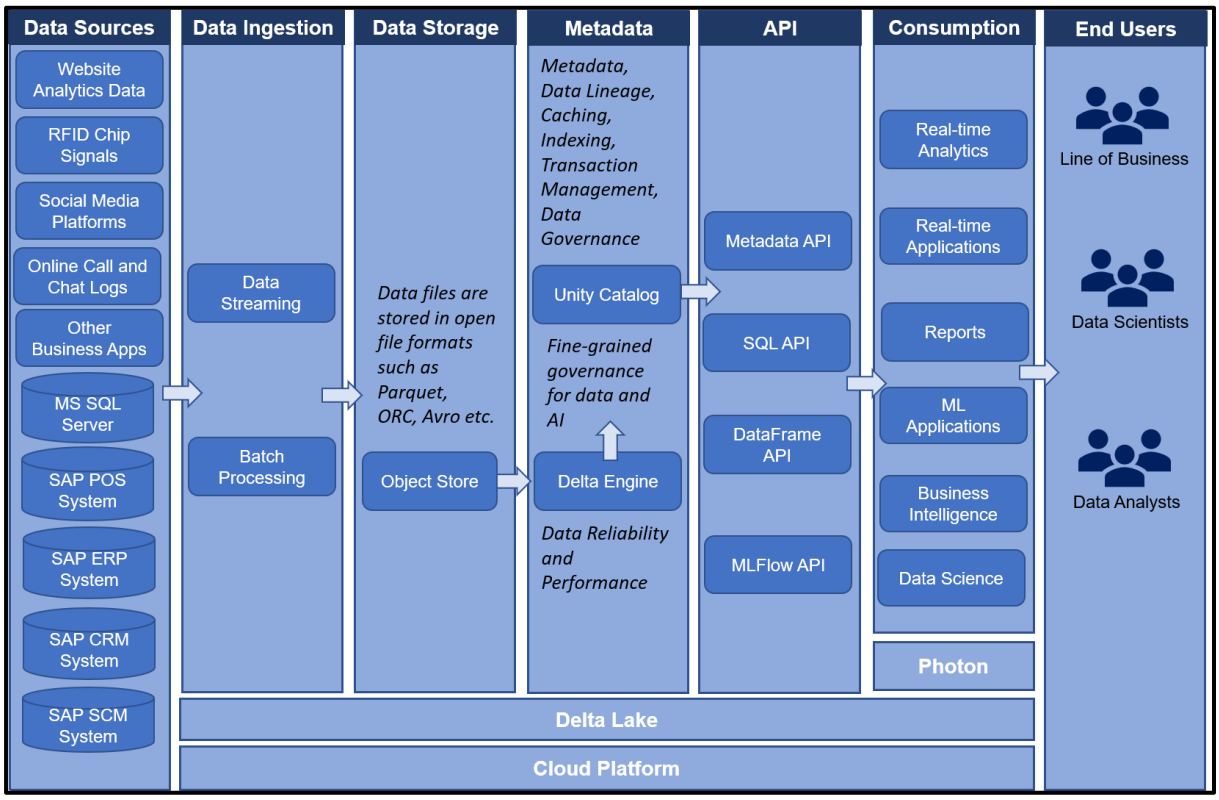

Figure 21.Databricks Data Lakehouse HLD(Alexsoft,2021;Databricks, no date;Phaujdar,2023)

As shown in Fig.21, the data flows from the data sources layer to the consumption layer. Since the data lakehouse layers were already covered in the previous section, the following sections will cover the key components of this architecture.

5.1.4.1. Delta Lake

Delta Lake is a storage layer that helps to replace the batch and streaming data in different formats from silos into a single data source(Databricks,no date a). It offers reliability, security and higher performance in the data lake(ibid). It is cost-effective, ACID-compliant and highly scalable(ibid). Fig.22 illustrates the components of Delta Lake.

Delta Lake Components	Functionality
Delta Sharing	It is an open protocol that enables companies to securely share data with other companies such as suppliers, auditors etc. irrespective of the data location(Databricks, no date a). Delta Lake leverages Apache Spark, enabling 48x times faster ETL workloads and 50x times faster time-to-insight(ibid).
Delta Live Tables	It automates data engineering in Delta Lake (Databricks, no date a). It enables the data engineering team to easily manage and develop ETL with declarative pipelines(ibid). It is proprietary to Databricks(Databricks,2023).
Delta Engine	It is a query optimizer for big data that uses Delta Lake(Databricks,2023). It can optimise the performance by pushing computation to the data in Spark SQL, Databricks SQL, and DataFrame operations(ibid).
Delta Logs	Delta Lake achieves atomicity by tracking all changes made by the users and storing them in Delta Logs(Databricks,2023).

Figure 22.Delta Lake Components

5.1.4.2. Unity Catalog

Unity Catalog is a centralised data governance tool based on SQL that resides on the top of Delta Engine in the data lakehouse(Databricks,no date d). It offers unified fine-grained auditing by maintaining an audit log of actions performed against the data and aids the company to meet its compliance and audit requirements(ibid). Fig.23 illustrates the key features of Unity Catalog in the data lakehouse.

Key Features of Unity Catalog	
Centralised governance for data and AI	Centrally govern data in files, tables, dashboards and ML models deployed in any cloud platform.
Built-in search and data discovery	Enables to search, understand and reference data across the data lake quickly. Offers secure data search by restricting search results based on the user access privileges
Performance and Scale	Offers low-latency metadata serving, resulting in improved query performance and auto-tuning of the tables.
Automated lineage for all workloads	Provides a unified and transparent view of the entire data ecosystem. Also, offers automated and granular lineage for all workloads in SQL, R, Python, Scala and across all asset types including tables, notebooks, workflows and dashboards. Lineage can be accessed using REST APIs as well as other existing APIs. Lineage graphs are restricted to users based on the access privilege to enforce security.
Integration with existing tools	Offers the flexibility to leverage the existing data catalogs and governance solutions to the data lakehouse, allowing the company to cut migration costs.
Secure data sharing	Unity Catalog can be securely shared with other companies such as partners, suppliers or auditors using Delta Sharing protocol even if they are not on Databricks, same cloud or on-premise irrespective of the data location.

Figure 23.Key Features of Unity Catalog(Databricks,no date d)

5.1.4.3. Photon Engine

Photon Engine is the next-gen engine on the Databricks data lakehouse platform that offers high-performance query executions at low cost(Databricks, no date c). It is

compatible with Apache Spark APIs and can be easily activated without modifying code and doesn't have any dependencies on the platform(ibid). It is vectorised native engine packaged along with Databricks Runtimes(DBR) and is written in C++ and designed to leverage the hardware power to offer high-performing queries(ibid).

5.1.4.4. Managed MLFlow

MLFlow is a lightweight API set and user interface(UI) that supports MLOps which enables automation of the ML lifecycle(Databricks,no date b). It is a managed service offered by Databricks that is built on open-source MLFlow and also provides security, reliability and scalability(ibid). Fig.24 and Fig.25 illustrate the key features and components of Managed MLFlow respectively.

Key Features of MLFlow	
Model Development	Automates ML lifecycle management through a standardized framework for production-ready ML model development. By leveraging MLFlow Recipes, the company can perform deliver high-quality models in production in rapid iterations.
Experiment Tracking	Automatically maintain the history of parameters, metrics, code and models from each experiment. Built-in integrations with Databricks Workspace and Notebooks help users to securely share, manage and perform the comparison of experiment results with respective artifacts and code versions.
Model Management	Enables model management easier by providing a central platform for discovering and sharing ML models, moving models from experimentation to testing and production environments, enabling integration of approval and governance workflows and CI/CD pipelines, and monitoring the ML model performance post-deployment. MLFlow Model Registry helps to share the knowledge.
Model Deployment	Allows quick production deployment for batch inference by leveraging the integrations such as Docker Containers, Azure ML or Amazon SageMaker. Additionally, it offers Databricks Job Scheduler to monitor production models and auto-managed clusters for scaling as per business requirements.

Figure 24.Key Features of Managed MLFlow(Databricks,no date b)

MLFlow Components	Functionality
MLFlow Tracking	Automatically maintain the record of parameters, metrics, code versions and artifacts for each experimental execution. Offers built-in tracking server without any extra configuration and securely manages experiments within the workspace by restricting access to users based on access permissions.
MLFlow Projects	Enables ML project management easier by allowing users to specify the software execution environment in the code configuration. Currently supports Conda, Docker container, and system environments.
MLFlow Models	Offers a standardized format for ML model packaging making it easily used by a wide range of downstream tools such as REST API real-time serving or Apache Spark batch inference. Additionally, it offers customized and built-in model flavors and allows quick deployment of models.
MLFlow Models Registry	Provides a central repository for registered MLFlow models. Each registered ML model has a unique name, version, stage, and other metadata. It keeps track of model versions automatically and assigns custom stages for each model version such as Production, Staging etc. Allows integration with CI/CD pipelines and automatically maintains model stage transitions.
MLFlow Recipes	Offers out-of-box connected components for building ML models and creates reusable iterations for reducing cost. Provides automated handoff from experimentation to production environments.

Figure 25.Managed MLFlow Components(Databricks,no date b)

6. Applications of the Proposed System in Joules

Having seen the architecture of the proposed system, Fig.26 depicts how the proposed system will address the existing challenges of Joules outlined in Fig.16.

Problem No.	Identified Root Cause	Solution
P1	High freight charges and wage costs resulted in increased product price(Provan and Wallace,2022).	Perform demand and trend forecasting using ML model to enhance operational efficiency, and reduce costs and wastage. Managed MLFlow will enable the technical team to easily integrate the model and enforce MLOps and thereby deliver an ML model that adds value to the business more quickly. Since the IT team is already following DevOps culture, they can quickly adopt the MLOps methodology.
P2	Improper inventory management, sales and fashion trend forecasting.	ML models will help to forecast fashion trends based on a diverse set of data such as market trends, sentiment analysis from social media, occasions, personal preferences etc. Data Lakehouse helps to perform real-time analytics and forecasting of the trends that enable the company to stock the estimated amount of the products in the store.
P3	Designers don't have proper insights into rapidly changing customer preferences and are unable to forecast the trends, resulting in deadstock.	Data Lakehouse will provide real-time analytics based on the data from diverse platforms such as social media, RFID chip signals from stores, website analytics and market trends, enabling designers to make quick data-driven design decisions. Photon engine will help to query data faster on the huge dataset and Unity Catalog will help to search relevant data quicker. The quicker insights will provide a competitive advantage in the market and help the company to release more innovative designs than competitors, resulting in higher sales revenue.
P4	Warehousing and shipping operations are not optimised. Also, the staff is not allocated properly as per the warehouse and delivery requirements.	Data Lakehouse can replace the data silos such as ERP, CRM,SCM and POS, integrate with relevant data more quickly and provides insights into the improvement areas that enable the company to achieve optimised operations and generate more sales and increase profit margins.

Figure 26.Solutions to the challenges faced by Joules.

7. Risks Involved and Mitigation Actions

Fig.27 depicts the risks involved in the implementation of the proposed system and mitigation actions.

Risk Category	Risk Description	Risk Level	Mitigation Action
Adoption	The adoption of BDA solutions might not add any value to the business if it is not implemented correctly without a clear understanding of the business requirements, expected outcomes and risks involved in it (Elia *et al.*,2020).	Moderate	A thorough analysis should be conducted before adopting the BDA solution to any business use case to ensure that it aligns with the business requirements and can produce expected outcomes(Elia *et al.*,2020). It should be carefully designed and implemented to avoid any undesired result(ibid). Furthermore, it should be continuously monitored and evaluated after adoption to check whether it is adding value to the business and to identify improvement areas(ibid).
Technical	Building Data pipelines involve technical complexity and requires technical expertise for implementation. Inefficiencies in the pipeline may result in inaccurate results and bottlenecks(Foley, 2022).	Moderate	Proper testing should be performed before deploying the proposed system to ensure that it is meeting the business objectives.
Technical	Data Lakehouse is relatively new architecture and it is not mature as data lake and data warehouse architectures(Kutay,2023).	Moderate	The technical team should keep up-to-date with the latest changes and patches in this architecture to overcome any potential challenges identified in the architecture and seek assistance from Databricks if required(Databricks, no date).
Implementation	Implementation of data lakehouse requires technical expertise and knowledge. (Databricks,no date;Wickramasinghe,2023)	Moderate	The company should train the staff on the new technology under consideration and employ technical experts if required. Also, the company can seek technical support from Databricks in the implementation and maintenance of the proposed system. Proper planning, training, analysis and testing before implementation can help to mitigate the implementation risks. (Databricks,no date;Wickramasinghe,2023)
Implementation	Data lakehouse stores raw data that can be unusable without proper security and catalog mechanisms (Wickramasinghe,2023)	High	Ensure proper implementation of Unity Catalog, data governance and security mechanisms for accuracy and better results.

Figure 27.Risks Involved and Mitigation Actions

8. Conclusion

The requirement of Joule's CEO was to propose a strategy that overcome the challenges faced by the company using BDA technology. As the initial step, this study analysed organisation structure, identified business objectives and challenges and conducted market research. This report also covered the benefits of using BDA in the clothing industry. Before proposing the new system, this study identified different data sources and performed an analysis of various data architectures such as data warehouse, data lake and data lakehouse. Data Lakehouse has been selected for the proposed system since it is capable of catering for the business requirements of Joules. Additionally, it analysed different vendors who offer data lakehouse platforms and Databricks was chosen as the suitable vendor for the company because they offer technical support and easy implementation without cloud platform dependency. The study covered the architecture of the proposed system in detail. Furthermore, it highlighted the benefits, challenges, and mitigation actions of the suggested approach.

9. References

Ahmad, S. *et al.* (2020) “Towards sustainable textile and apparel industry: Exploring the role of business intelligence systems in the ERA of Industry 4.0,” *Sustainability*, 12(7), pp. 1–23. Available at: https://doi.org/10.3390/su12072632.

AlexSoft (2021) *Data Lakehouse: Concept, key features, and architecture layers*, *AltexSoft*. AltexSoft. Available at: https://www.altexsoft.com/blog/data-lakehouse/ (Accessed: May 1, 2023).

Balasubramaniam, L. (2022) *Implementing Data Lakehouse in an enterprise*, *Medium*. Medium. Available at: https://medium.com/@lackshub/implementing-data-lakehouse-in-an-enterprise-1cb559df714d (Accessed: May 1, 2023).

BuiltWith (no date) *Joules.com technology profile on BuiltWith*, *BuiltWith*. Available at: https://builtwith.com/?https%3A%2F%2Fwww.joules.com%2F (Accessed: April 30, 2023).

Bython (2023) *Understanding the benefits of customer segmentation*, *Bython*. Available at: https://bython.com/benefits-of-customer-segmentation/ (Accessed: April 29, 2023).

Certilogo (2022) *The role of data analytics in Fashion*, *certilogo*. Available at: https://discover.certilogo.com/blogs/insights/role-of-data-analytics-in-fashion#:~:text=Data%20can%20help%20brands%20better,sales%20and%20improve%20customer%20retention.&text=Brands%20can%20understand%20customer%20needs,thus%20providing%20tailored%20customer%20service. (Accessed: April 29, 2023).

Ceron, T.S. (2023) *UK clothing retailing market report - market size & growth*, *Mintel Store*. Available at: https://reports.mintel.com/display/1100483/?fromSearch=%3Ffreetext%3Dclothing%26resultPosition%3D1 (Accessed: April 30, 2023).

Clark, H. (2023) *10 best customer segmentation tools in 2023*, *The CX Lead*. Available at: https://thecxlead.com/tools/best-customer-segmentation-tools/ (Accessed: April 30, 2023).

Databricks (no date) *Data Lakehouse platform by Databricks - schedule a demo now!*, *Databricks*. Available at: https://www.databricks.com/product/data-lakehouse?_gl=1%2Augg1ia%2A_gcl_aw%2AR0NMLjE2ODI5MjQ0MzIuQ2p3S0NBand4cjJpQmhCSkVpd0FkWEVDdzhPdVJuUHZzV3dRUVNfSU1UMTc2TEIwaDUwNEEwS1hvel8wTTZXX2tMTk9iRDhLcG5McF9ob0MxeWdRQXZEX0J3RQ.. (Accessed: May 1, 2023).

Databricks (no date a) *Delta Lake on databricks - schedule a demo now!*, *Databricks*. Available at: https://www.databricks.com/product/delta-lake-on-databricks (Accessed: May 1, 2023).

Databricks (no date b) *Managed mlflow*, *Databricks*. Available at: https://www.databricks.com/product/managed-mlflow (Accessed: May 1, 2023).

Databricks (no date c) *Photon*, *Databricks*. Available at: https://www.databricks.com/product/photon (Accessed: May 1, 2023).

Databricks (no date d) *Unity catalog*, *Databricks*. Available at: https://www.databricks.com/product/unity-catalog (Accessed: May 1, 2023).

Databricks (2023) *What are all the delta things in Databricks?*, *Databricks on AWS*. Available at: https://docs.databricks.com/introduction/delta-comparison.html (Accessed: May 1, 2023).

Elia, G. *et al.* (2020) "A multi-dimension framework for value creation through Big Data," *Industrial Marketing Management*, 90, pp. 617–632. Available at: https://doi.org/10.1016/j.indmarman.2020.03.015.

Foley, T. (2022) *What is a data pipeline and how to build one*, *2nd Watch*. 2nd Watch. Available at: https://www.2ndwatch.com/blog/what-is-a-data-pipeline-and-how-to-build-one/ (Accessed: April 30, 2023).

Hickins, M. (2023) *How data analytics informs the Fashion Industry*, *Oracle*. Available at: https://www.oracle.com/retail/fashion/fashion-analytics/#:~:text=In%20fashion%2C%20data%20analytics%20can,accurately%20so%20that%20they%20can (Accessed: April 27, 2023).

Jolly, J. (2022) *Joules to appoint administrators as rescue talks fail*, *The Guardian*. Guardian News and Media. Available at: https://www.theguardian.com/business/2022/nov/14/joules-to-appoint-administrators-fashion-retailer-stores#:~:text=Joules%2C%20which%20is%20best%20known,for%20its%20posh%20winter%20wellies. (Accessed: April 28, 2023).

Joules (no date) *Recycle*, *Joules*. Available at: https://www.joules.com/Recycle (Accessed: April 28, 2023).

Kava , P. and Gong, C. (2021) *Blogs*, *Amazon*. Greenhaven Press/Gale. Available at: https://aws.amazon.com/blogs/big-data/build-a-lake-house-architecture-on-aws/ (Accessed: May 1, 2023).

Kiron, M.I. (2021) *Big Data in fashion industry: Impact, benefits and application*, *Textile Learner*. Available at: https://textilelearner.net/big-data-in-fashion-industry/ (Accessed: April 29, 2023).

Kossendey, R. (2023) *Lakehouse - Databricks vs. AWS EMR*, *Medium*. claimsforce. Available at: https://medium.com/claimsforce/lakehouse-databricks-vs-aws-emr-87e30c00b791 (Accessed: May 1, 2023).

Kutay, J. (2023) *Data Warehouse vs. Data Lake vs. Data Lakehouse: An overview of three cloud data storage patterns*, *Striim*. Available at: https://www.striim.com/blog/data-warehouse-vs-data-lake-vs-data-lakehouse-an-overview/ (Accessed: May 1, 2023).

Lamba, K. and Singh, S.P. (2017) “Big Data in operations and Supply Chain Management: Current trends and future perspectives,” *Production Planning & Control*, 28(11-12), pp. 877–890. Available at: https://doi.org/10.1080/09537287.2017.1336787.

Lin, R.-H. *et al.* (2021) “Applied Big Data Analysis to build customer product recommendation model,” *Sustainability*, 13(9), p. 4985. Available at: https://doi.org/10.3390/su13094985.

Logunova, I. (2022) *Data Warehouse vs. Data Lake vs. Data Lakehouse: Key Pros & Cons*, *Serokell Software Development Company*. Serokell. Available at: https://serokell.io/blog/data-warehouse-vs-lake-vs-lakehouse (Accessed: May 1, 2023).

Olabiyi, K. (2022) *Fashion and big data*, *Medium*. DATA4FASHION. Available at: https://medium.com/data4fashion/fashion-and-big-data-application-cb946dd76844 (Accessed: April 28, 2023).

Phaujdar, A. (2023) *What is a big data pipeline? Key Components, Architecture & Use Cases - learn*, *Hevo*. Available at: https://hevodata.com/learn/big-data-pipeline/ (Accessed: May 1, 2023).

Provan, S. and Wallace, A. (2022) *Joules set to enter administration after failing to secure fresh funding*, *Subscribe to read | Financial Times*. Financial Times. Available at: https://www.ft.com/content/a8dc1f09-6318-4b3b-940e-51324c37f38b (Accessed: April 28, 2023).

Rehan (2022) *Data capture: What is data capture and why is it so crucial?*, *AmyGB.ai - Intelligent Document Processing Company*. Available at: https://www.amygb.ai/blog/what-is-data-capture#:~:text=The%20Benefits%20of%20Using%20Data%20Capture&text=Automated%20data%20collection%20excels%20above,the%20simplified%20data%20extraction%20process. (Accessed: April 27, 2023).

SAP (no date) *Fashion and apparel industry software and ERP*, *SAP*. Available at: https://www.sap.com/uk/industries/fashion-apparel.html (Accessed: April 30, 2023).

SEKO (no date) *Joules case study: Seko Logistics*, *Global Logistics Freight & Delivery Experts | SEKO Logistics*. Available at:

https://www.sekologistics.com/uk/case-studies/joules/ (Accessed: April 28, 2023).

SimilarWeb (2023) *Joules.com competitors - top sites like joules.com | similarweb*, *SimilarWeb*. Available at: https://www.similarweb.com/website/joules.com/competitors/ (Accessed: April 30, 2023).

Wickramasinghe, S. (2023) *Data lakehouses: Everything you need to know*, *Splunk*. Available at: https://www.splunk.com/en_us/blog/learn/data-lakehouses.html#:~:text=A%20Data%20Lakehouse%20is%20also,benefit%20organizations%20in%20different%20ways. (Accessed: May 1, 2023).

Zavialova, S. (2023) *Fashion eCommerce report 2023*, *Statista*. Available at: https://0-www-statista-com.pugwash.lib.warwick.ac.uk/study/38340/ecommerce-report-fashion/ (Accessed: April 30, 2023).